DOCUMENT RESUME

ED 273 673 TM 860 525

AUTHOR Carlson, Robert E.
TITLE Field-Test Data vs. Real-Test Data.
PUB DATE Apr 86
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - General (140)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cutting Scores; Graduation Requirements; High Schools; *Item Analysis; Minimum Competency Testing; Occupational Tests; Performance Factors; *Pretesting; *Response Style (Tests); Sampling; Teacher Certification; *Test Construction; Testing Problems
IDENTIFIERS Alabama High School Graduation Examination; Georgia Teacher Certification Testing Program

ABSTRACT

Field tests of test items are useful in developing cutting scores for important tests, such as teacher certification examinations and high school graduation tests, because they indicate test item quality and estimate future examinee performance. However, field test data may provide a faulty indication of examinee performance on the real test. Examinee performance is often better when the test counts than it is on the field test. Based on the field test for Texas's high school graduation test, it was estimated that about 25 percent of the examinees would fail both the English and mathematics tests. However, only nine percent failed the English test and twelve percent failed mathematics. For the Alabama High School Graduation Test, actual performance exceeded field test performance in 53 of 54 competency areas. With the Georgia Teacher Certification Tests, item p-values typically increased ten or more points from field test to real test. Actual failure rates on the Arkansas educator recertification test were also much lower than those of field test examinees. Factors which contribute to these differences include the representativeness of the field test sample, examinee motivation, examinees' preparation for actual tests, and examinees' ability to repeat actual tests. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
RE Carlson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

FIELD-TEST DATA VS. REAL-TEST DATA*

Robert E. Carlson

IOX Assessment Associates

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED273673

For high-stakes tests, such as those used for educator certification/recertification and high school graduation, the determination of cut-scores is of paramount importance. The effects on examinees who do not achieve the specified cut-score on such tests can be devastating. It is imperative, therefore, that those responsible for setting cut-scores do so after much thought and careful deliberation of all relevant information available. One important source of information is a field test of test items.

The Use of Field-Test Data When Setting Cut-Scores

Field tests are a common component of test development. They can provide both statistical indices of item quality as well as data useful in estimating future examinee performance. It is generally thought that the way in which field-test examinees perform can give some indication of how future examinees will perform. For example, by using Rasch analyses after an actual test form has been created, it is possible to develop a table showing, for each score point, what percentage of the field-test examinees would be expected to have met or exceeded that score had they taken the real test form. This is useful when trying to estimate how many examinees will pass the test at each possible

TM860 525

* Presented at the annual meeting of the American Educational Research Association, San Francisco, California, April 16-20, 1986.

cut-score. A major concern of standard-setters is the impact that a particular cut-score will have on the population of examinees. They are concerned about how many individuals will fail at the selected cut-score.

Thus, field-test data is a valuable type of information for standard-setters. This type of data is often considered with data from other sources by those who must determine a cut-score. It is not clear, however, how accurate field-test data are as indicators of future examinee test performance. That is, standard-setters may be basing their decisions on field-test performance data that are not representative of the future performance data on the actual tests to which the cut-score will be applied. Field-test data may, in fact, be providing a faulty indication of how examinees will do on the real test.

I am in the midst of a study, the purpose of which is to investigate the accuracy of field-test data as indicators of future examinee test performance. This morning, I would like to share with you some of my preliminary findings and thoughts about this issue.

The Accuracy of Field-Test Data

I have collected some anecdotal and empirical evidence indicating that, on high-stakes tests, examinee performance when the test counts is better than it is on the field test. For example, based on the field test of the Texas high school graduation test, it was estimated that about 25% of the examinees would fail the English language arts test and an equal number would fail the mathmatics test. When the actual test was administered for the first time, however, only about 9% failed the English

language arts test and about 12% failed the mathematics test (personal communication, Texas Education Agency).

A comparison of field-test and real-test p-values of items on the Alabama High School Graduation Test shows that real-test examinees performed at a higher level (Steele, 1985). The increase in average p-values for items on the reading test was about five points, on the math test about nine points, and on the language test about six points. A comparison of average p-values of items in each of 54 competency areas tested on the examination shows that, in all but one competency area, performance on the real test exceeded performance on the field test. For 14 competencies, the average p-value increase was 10 points or more.

On the Georgia Teacher Certification Tests, a state department official told me that item p-values typically increased 10 or more points from field test to real test (personal communication, Georgia Department of Education). On the Arkansas educator recertification test, the actual failure rates were much less than those based on the field-test examinees. In reading, 26% of the field-test examinees would have failed, yet only 5% actually failed. In mathematics, 13% of the field-test examinees would have failed, but only 3% actually did. In writing, the failure rate went from 9% to approximately 4%.

Thus, there seems to be an indication that field-test examinee performance differs from real-test examinee performance. For administrators using field-test data as one source of information when determining passing standards, this is an important issue. What is it that might explain such differences?

Factors Contributing to Differences Between Field-Test and Real-Test Performance

Sampling

An obvious cause of differences between field-test and real-test performance is that the field-test sample is not representative of the population of examinees who subsequently take the real test. Obtaining a truly representative sample, however, is problematic. New items in an already implemented testing program can be field-tested by embedding them in operational test forms. For the initial development of a test, however, this option is not available. Two key sampling decisions have to be made: from what population to sample and how to select individuals from the population.

For high school graduation tests, the difficult issue is the decision regarding the population from which to sample. Most choices either involve sampling from a different population (e.g., sampling from a class of students that is not the class that will actually be tested) or sampling from the class to be tested but field-testing at a different point in the school year than when the real test will be administered (e.g., field-testing tenth-grade students in the fall on a test they will take in the spring). In the first case, it is likely that there are differences between the field-test population and the target population that would make the field-test data suspect for use in projecting performance levels on the real test. In the second case, the students will have had six or seven more months of instruction when they take the real test than when they took the field test. This will surely create differences in performance.

Similar difficulties are encountered when determining a field-test sample for an initial certification test for educators. One strategy, recommended by Berk (1986), is to sample from two populations: successful, practicing teachers and students early in their teacher-training programs. Neither is the population that will actually be tested, that is, teacher candidates who have completed teacher education programs; but, by using performance data from both groups, standard setters can, in a sense, interpolate an estimate of examinee performance on the real test. New problems emerge, however. Unlike high school students, inservice teachers and teacher education students are not "captive groups." It may be difficult to obtain the cooperation of these groups, and teacher education faculty members may be reluctant to mandate the test or give up class time to administer it. Practicing teachers may not have the time or interest to participate.

Turning to educator recertification tests, the tricky issue is deciding how to recruit a representative sample for the field test. Unlike with student tests, it is possible to select participants for a field test of an educator recertification test from the population that will eventually be tested, that is, the population of educators required to take the test. Unlike student tests, however, it can be difficult to obtain a representative sample to actually participate in the field test. The field test can rarely be made mandatory. The only real incentive for an educator to participate in a field test is the opportunity to preview potential test items. This might attract only those educators who are not confident of their ability to do well on

the test. Thus, less able educators might agree to participate in the field test at higher rates than others. The political environment in which an educator recertification test is being developed can also affect the chances of obtaining a representative sample.

These problems were illustrated in Arkansas (Collins, 1986). Approximately 800 educators were randomly selected and invited by mail to participate in the field test of the state's educator recertification examination. The acceptance rate was so low that more educators had to be selected and invited. Eventually, 1500 were invited before 600 agreed to participate. On the day of the field test, only 498 of the 600 actually attended. As mentioned earlier, field-test performance in Arkansas greatly underestimated real test performance--clearly, the problems in obtaining a representative sample were a source of the difference.

Motivation

Once the sampling dilemma is dealt with, other problems surface that may very well contribute to differences between field-test and real-test performance. One major factor is that of motivation. It is likely that, if examinees know the test they are taking doesn't count, they will not try as hard as they would if they knew the test counted. With high-stakes tests for adults and older students, field-test examinees typically know it doesn't count. This is the factor mentioned to me most often by state level test development people when talking about the accuracy of field-test data. When people know that their high

school diplomas or their licenses are on the line, they naturally put forth a good deal of effort.

High school students asked to participate in a field test of a graduation test are generally obliged to comply; they are somewhat of a captive sample. They might be interested in seeing the test, but knowing that "the test has no teeth" is likely to affect their motivation to do well. Future or inservice educators participating in a field test of a certification or recertification exam are probably more interested in seeing potential test items than in performing their best.

Level of motivation to do well on a test involves a number of behaviors, but the key variable is probably the degree of effort expended. This variable may come into play most often with relatively difficult items. There is little incentive for a field-test examinee to spend the extra time needed to solve a particularly vexing test item. On the real test, however, there is such an incentive. One possible indication of differences in level of effort comes from the experience with the Texas educator recertification test. Most field-test examinees completed the field test in two or three hours. On the real test, however, many examinees took much longer (personal communication, Texas Education Agency).

Amount of Preparation

Another potential cause of differences between field-test and real-test performance, related to motivation, is the amount of time examinees spend preparing for the test. Field-test examinees are not likely to have received any descriptive material about the test before they participate in the field

test. If they do receive such information, there is little incentive to spend much time reviewing, studying, or practicing the skills and content to be tested. For examinees about to take the real test, however, the situation is quite different. Test descriptions and study guides are often disseminated, and preparation courses are offered. Knowing that the test will count provides the incentive for making use of such resources to prepare for the test.

While this factor seems most prevalent in educator tests, it also arises in connection with high school graduation tests. Differences in preparation can result from field-test sampling strategies as discussed above. In addition, although it doesn't seem very likely that high school students taking the test for the first time would spend non-school time preparing for the test, it is likely that teachers, after knowing what skills are tested, will spend more time emphasizing these skills. Preparing for a test probably has a differential effect across skills. For example, certain mathematics and even language arts skills may be more amenable to review and practice than some of the higher level reading skills.

Multiple Opportunities to Pass

Finally, when using field-test data to make estimates of how many real-test examinees might fail at each cut-off score, it is important to recognize that such estimates typically assume a single test administration. For all high-stakes tests of which I am aware, however, examinees are given more than one opportunity to pass. Thus, what really should be estimated is the cumulative pass rate for the test, that is, the percentage of examinees who

eventually pass the test given the number of opportunities allowed (Solomon, 1986). Examinees who fail a test the first time are often given score reports that indicate areas of strength and weakness. Thus, these examinees' preparation efforts for subsequent administrations of the test can be more focused. Remediation programs and materials are often made available to those who must retake the test.

Focusing on cumulative pass rates often provides a very different estimate of the impact of a testing program than an estimate based on only one administration. On the Texas educator recertification test, for instance, field-test data showed that about 12% of the field-test examinees would have failed one or both parts of the test. After considering that educators have two chances to pass and that remediation efforts are likely to be intense after the first administration, the estimate of how many educators would not pass the test was revised to about 5%. In the Georgia Teacher Certification Program, almost 22% of the first-time examinees failed. These examinees were tracked through three more test administrations. After the fourth administration, only about 13% still had not passed (Solomon, 1986). Clearly, then, those responsible for setting cut-scores need to consider the cumulative pass rate when going about their task.

Summary

There are a variety of factors that might cause differences in examinee performance from field test to real test on high-stakes tests. It is important that those who are responsible for

determining cut-scores recognize that field-test data are apt to be only moderately accurate in predicting real-test examinee performance. I have presented some preliminary data that indicate that field-test performance underestimates real-test performance. In my study, I will be collecting both field-test and real-test data from a variety of tests and will see if this preliminary finding persists. If my early findings are confirmed, this clearly does not mean that field-test data are not useful in setting passing standards. Some performance data are better than no data at all. But standard-setters must realize that field-test data provide only a rough estimate (most likely an underestimate) of how examinees will actually perform on the real test.

References

Collins, J.L. (1986, April). Procedural pitfalls in educator recertification testing: An Arkansas retrospective. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Berk, R.A. (1986). Judgmental and statistical item analyses of teacher certification tests. In W.P. Gorth & M.L. Chernoff (Eds.), Testing for teacher certification (pp.165-183). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Solomon, L.M. (1986). Analysis of candidates who retake the teacher tests. In W.P. Gorth & M.L. Chernoff (Eds.), Testing for teacher certification (pp.89-98). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Steele, D.J. (1985, November). A descriptive comparison of test items utilized in pilot and live administrations of the Alabama High School Graduation Examination. Paper presented at the meeting of the Mid-South Educational Research Association, New Orleans, LA.